DOCUMENT RESUME

ED 434 121 TM 030 082

AUTHOR Thompson, Russel L.
TITLE Reliability Generalization: An Important Meta-Analytic Method, Because It Is Incorrect To Say, "The Test Is Unreliable."
PUB DATE 1999-01-21
NOTE 13p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, January 21-23, 1999).
PUB TYPE Information Analyses (070) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Meta Analysis; *Reliability; Research Methodology; Scores
IDENTIFIERS *Generality

ABSTRACT

Many researchers fail to understand that reliability is a function of scores, not tests. This paper provides an explanation of the distinction as well as a description of the reliability generalization meta-analysis technique. Reliability generalization meta-analysis can provide a way to aggregate test score reliability coefficients from prior studies based on the characteristics of those studies. The resulting information can help researchers anticipate score reliability and identify characteristics for improving score reliability. (Contains 17 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 434 121

Running head: RELIABILITY GENERALIZATION

Reliability Generalization: An Important Meta-Analytic Method, Because It Is Incorrect To Say, "The Test Is Unreliable"

Russel L. Thompson

Texas A&M University, 77843-4225

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Russel Thompson
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

TM030082

Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, January 21, 1999.

BEST COPY AVAILABLE

Abstract

Many researchers fail to understand that reliability is a function of scores, not tests. This paper provides an explanation of the distinction as well as a description of the reliability generalization meta-analysis technique. Reliability generalization meta-analysis can provide a way to aggregate test score reliability coefficients from prior studies, based on the characteristics of those studies. The resulting information can help researchers anticipate score reliability and identify characteristics for improving score reliability.

It is unfortunately all too common to find authors of education and psychology journal articles describing the "reliability of the test" or stating that "the test is reliable." Such statements fail to recognize that reliability is a characteristic of scores, and not of tests. As {4 /id Pedhazur & Schmelkin 1991} noted, "Statements about the reliability of a measure are … [inherently] inappropriate and potentially misleading" (p. 82).

Similarly, Gronlund & Linn (1990) emphasized that the "results" are reliable, rather than "an evaluation instrument." They wrote,

> Reliability refers to the results obtained with an evaluation instrument not to the instrument itself… Thus, it is more appropriate to speak of the reliability of the "test scores" or the "measurement" than of the "test" or the "instrument" (p. 78, emphasis in original).

Rowley (1976) noted, "It needs to be established that an instrument itself is neither reliable or unreliable… A single instrument can produce scores which are reliable, and other scores which are unreliable" (p. 53, emphasis added). To summarize, it must be clearly understood that a

test is not 'reliable' or 'unreliable'. Rather, "reliability is a property of the scores on a test for a particular group of examinees" Crocker & Algina (1986) p. 144, emphasis added.

Because tests are not reliable *per se*, this means that score reliability fluctuates from study to study, and must be investigated in each study. The purpose of the present paper is to explain an innovative new method for evaluating the sources of score measurement error variances as these occur across studies: the *reliability generalization* method (Vacha-Haase, 1998).

Reliability generalization is an extension of the notable method, validity generalization, described by Schmidt & Hunter (1977) and Hunter & Schmidt (1990). In validity generalization inquiries (Schmidt & Hunter, 1977), studies are used as the unit of analysis, and means, standard deviations and other descriptive statistics are computed for the validity coefficients across studies. The validity coefficients across studies may also be used as the dependent variables in regression or other analyses. In these analyses, the features of the studies (e.g., sample size, types of samples, ages of participants) that best predict the variations in the obtained validity coefficients are investigated.

The same thing can be done to investigate reliability coefficients for a given measure across studies, as proposed by Vacha-Haase (1998). The method can be used to characterize for a given test (a) the typical reliability of scores across studies, (b) the amount of variability in reliability coefficients, and (c) the sources of variability in reliability coefficients across studies. The present paper provides an accessible summary of Vacha-Haase's important reliability generalization method.

The reliability generalization process initially requires the researcher to identify all prior studies that report reliability coefficients for the test under investigation. Studies must use the same methods for measuring reliability.

Huck & Cormier (1996) list three classical methods for measuring internal consistency reliability: split-half reliability coefficient, Kuder-Richardson #20, (also known as KR-20), and Cronbach's alpha. Of course, even more sophisticated "modern" (i.e., non-classical reliability coefficients can also be computed, such as Generalizability Coefficients (cf. Eason, 1991; Thompson, 1991)). Huck & Cormier (1996) emphasize that reliability estimates do not necessarily generalize across methods, so it is important

to identify the types of reliability statistics that will be used in the reliability generalization study.

The next step in the reliability generalization process is to identify common information that is provided in each study (e.g., sample size, gender, age), as well as natural subscore divisions that may be reported appropriately for the test. These data are then coded and each piece of information becomes a dependent variable for the statistical analysis of the reliability scores of the test.

Vacha-Haase (1998) demonstrates statistical treatments of reliability coefficients that can be applied as a part of the Reliability Generalization analysis. Descriptive statistics can be computed to describe central tendency and variability of reliability coefficients across studies. These statistics can give researchers a benchmark to compare reliability coefficients for scores in their study. Further statistical analyses can be conducted to discover which variables (e.g., sample size, type of reliability coefficient, characteristics of study participants) contribute most to, (or detract most from), test score reliability.

Examples of Reliability Estimates

The following excerpts from prior studies of the NEO-PI-R (Costa & McCrae, 1992), a measure of the five-factor model of personality, will demonstrate the language that is typically used to describe reliability data. These excerpts will also demonstrate some decisions that the researcher may need to make in collecting the reliability generalization data. In addition, examples of possible independent variables are given that a researcher might choose to include in a reliability generalization analysis.

Some studies report reliability estimates from prior research rather than reliability estimates for the data in the current study. For example, McCrae (1987) wrote "Internal consistency and 6-month retest reliability for the Neuroticism, Extraversion, and Openness scales range from .85 to .93" (p. 1260). In general, this practice can be identified by the reference to prior research. Studies that are reporting reliability estimates for their data will often precede their reliability estimates with a phrase such as "In the present study".

Another way researchers support the reliability of their study without actually calculating estimates for their data is to make general statements about the reliability. For example, MacDonald, Anderson, Tsagarakis,

& Holland (1994) wrote: "a fair amount of research has been done and excellent support for the validity and reliability of the domains has been consistently reported" (p. 341). Neither of these approaches provides any information about the reliability of the data in a particular study. These studies cannot be included in the reliability generalization meta-analysis.

Another unusable form of reporting of reliability estimates is seen in (Lay, 1997) in which it was reported that "Cronbach's alpha coefficients across the three separate samples ranged from 0.85 to 0.90" (p. 271). Since it is unclear which reliability estimate should go with which variable, this study would be rejected for reliability analysis.

The following excerpts are examples of studies that do report usable reliability estimates. All of the reliability estimates in the following examples are in the form of Cronbach's alphas. Judge, Martocchio, & Thoresen (1997) reported usable reliability data in the form of Cronbach's alphas: "Coefficient alphas for the personality scales were as follows: Neuroticism, α = .91; Extraversion, α = .87; Openness to Experience, α = .92; Agreeableness, α = .82; and Conscientiousness, α = .88" (p. 751). Cellar, Miller, Doverspike, & Klawsky (1996) reported that "The internal

consistency reliabilities for the five factor scales calculated for our sample were .84 (n=359), .72 (n=359), .71 (n=359), .78 (n=359), and .85 (n=362) for Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness, respectively" (p. 699). Note that the number of subjects is different for each of the five factors. The reliability generalization researcher could account for this by encoding a separate n for each subscale. Costa & McCrae (1995) reported "In the present sample, internal consistencies for the five domains were .92, .89, .89, .87, and .91 for N, E, O, A, and C, respectively" (p. 312).

Some examples from the prior studies that could have been encoded and included as independent variables in the reliability generalization are: number of subjects, mean, and standard deviation for the test scores. Participant characteristics that could have been included were age mean, median, range, standard deviation; gender; marital status; race/ethnicity; education level; number of children; and retirement status. Of course, the nature of the test and the type of study using the test will largely determine the dependent variables that are available for inclusion in the analysis. Once the data are encoded, statistical analyses (e.g., regression) can be conducted to

determine the influence of the encoded independent variables on test score reliability.

Summary

This paper has discussed the value of learning more about the psychometric properties of test scores through meta-analysis of score reliability across multiple studies. The examples of usable and unusable reports of reliability estimates, as well as language used to identify them, should provide a guide for researchers conducting reliability generalization studies. The results of a reliability generalization study will provide researchers with a better understanding of the reliability of scores obtained for tests in their particular studies as well as test characteristics that contribute most to score reliability in future studies.

References

Cellar, D.F., Miller, M.L., Doverspike, D.D., & Klawsky, J.D. (1996). Comparison of factor structures and criterion-related validity coefficients for two measures of personality based on the five factor model. Journal of Applied Psychology, 81(6), 694-704.

Costa, P.T., & McCrae, R.R. (1992). Revised NEO Personality Inventory and NEO Five Factor Inventory Professional Manual. Odessa, FL: Psychological Assessment Resources.

Costa, P.T., & McCrae, R.R. (1995). Primary traits of Eysenck's P-E-N system: Three- and five-factor solutions. Journal of Personality & Social Psychology, 69(2), 308-317.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Eason, S. (1991). Why generalizability theory yields better results than classical test theory: A primer with concrete examples. In B. Thompson (Ed.), Advances in educational research: Substantive findings, methodological developments. (pp. 83-98). Greenwich, CT: JAI Press.

Gronlund, N.E., & Linn, R.L. (1990). Measurement and evaluation in teaching. (6th ed.). New York: Macmillan.

Huck, S.W., & Cormier, W.H. (1996). Reading statistics and research. (2nd ed.). New York: Harper Collins College Publishers.

Hunter, J.E., & Schmidt, F.L. (1990). Methods of meta-analysis: Correcting error and bias in research findings. Newbury Park, CA: Sage.

Judge, T.A., Martocchio, J.J., & Thoresen, C.J. (1997). Five-Factor model of personality and employee absence. Journal of Applied Psychology, 82(5), 745-755.

Lay, C.H. (1997). Explaining lower-order trats through higher-order factors: the case of trait procrastination, conscientiousness, and the specificity dilemma. European Journal of Personality, 11 , 267-278.

BEST COPY AVAILABLE

MacDonald, D.A., Anderson, P.E., Tsagarakis, C.I., & Holland, C.J. (1994). Examinations of the relationship between the Myers-Briggs type indicator and the NEO personality inventory. Psychological Reports, 74, 339-344.

McCrae, R.R. (1987). Creativity, divergent thinking, and openness to experience. Journal of Personality & Social Psychology, 52(6), 1258-1265.

Pedhazur, E.J., & Schmelkin, L.P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Erlbaum.

Rowley, G.L. (1976). The reliability of observational measures. American Educational Research Journal, 13, 51-59.

Schmidt, F.L., & Hunter, J.E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.

Thompson, B. (1991). Review of Generalizability theory: A primer by R.J. Shavelson & N.W. Webb. Educational and Psychological Measurement, 51, 1069-1075.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. Educational and Psychological Measurement, 58(6), 20

BEST COPY AVAILABLE